Cell, Vol. 112, 151–155, January 24, 2003, Copyright ©2003 by Cell Press

Letter to the Editor

The ATRs, ATMs, and TORs Are Giant HEAT Repeat Proteins

We show a generally applicable protein sequence analysis that describes the non-kinase portions of the ATRs, ATMs, and TORs as giant helical (HEAT) repeat domains that differ from one another by the addition, subtraction, and rearrangement of specific HEAT repeat units.

The PIK-like protein superfamily comprises a functionally diverse set of molecules united by a common C-terminal PI3K-like kinase domain. Of the six identified PIK-like subfamilies, five are implicated in signal transduction. ATRs (*AT*M and *R*ad3 related) and ATMs (*a*taxia *t*elangiectasia *m*utated) respond to chromosome status in both unchallenged and damaged cells (Shiloh and Kastan, 2001; Abraham, 2001; Cha and Kleckner, 2002); TORs (*t*arget *o*f *r*apamycin) modulate protein translation in accord with nutrient levels (Schmeizle and Hall, 2000); DNA-PKs (*DNA*-dependent *p*rotein *k*inase) monitor nonhomologous endjoining (Ma et al., 2002); and SMG-1s perform mRNA surveillance (Yamashita et al., 2001). Finally, the TRRAPs (*Tr*ansformation/*Tr*anscription Domain *A*ssociated *P*rotein) retain PI3K homology but lack the ATP binding residues essential for kinase activity (Vassiley et al., 1998; McMahon et al., 1998; Grant et al., 1998). All six subfamilies are believed to be tumor suppressors (Rosen et al., 2000; McMahon et al., 1998; Shiloh and Kastan, 2001).

All PIK-like proteins are quite large, 270–450 kDa, with the kinase domain accounting for only 5%–10% of total sequence. Relatively little is known about the nature or specific function(s) of the remainder of these proteins, in part because only sparse amino acid sequence homology has been detected among and sometimes even within (e.g., ATRs, ATMs) subfamilies. However, the first half of the N-terminal region of the TORs is composed of HEAT (*h*untingtin, *e*longation factor 3, *A* subunit of protein phosphatase 2A and *T*OR1) repeats (Andrade and Bork, 1995). A single HEAT repeat unit is a pair of interacting anti-parallel helices linked by a flexible "intra-unit" loop (Figure 1A). HEAT repeats occur in series, with adjacent units linked by flexible inter-unit loops (Andrade et al., 2001a, 2001b). In crystallographically analyzed proteins, HEAT repeat domains form superhelical scaffolding matrices, often when engaged with other macromolecules (Cingolani et al., 1999; Chook and Blobel, 1999; Groves et al., 1999). Secondary degrees structure prediction algorithms (e.g., Hierarchical Neural Network [HNN]; http://us.expasy.org/tools) find that all PIK-like proteins have high levels of α helicity, $\geq$50% (Figure 1B, data not shown). We therefore examined the possibility that all of these proteins could be comprised largely of helical repeat units. We describe an analysis of the ATR, ATM, and TOR PIK-like subfamilies, which reveals that the non-kinase portions of these proteins are composed almost entirely of HEAT repeats; evolutionary relationships within and among subfamilies emerge. These particular findings very likely extend to other PIK-like proteins, and the described approach should be generally applicable to identification and analysis of other HEAT repeat proteins.

We began our analysis by constructing alignments of the protein sequences for each subfamily. The TORs are sufficiently similar that they can be well-aligned with each other using standard alignment algorithms. However, the ATRs and ATMs are each sufficiently divergent that no satisfactory comprehensive alignment emerges from such an approach. We therefore examined each subfamily by generating a large number of alignments from overlapping subsets of sequences and subsequently reconciling these alignments with one another. For each sequence within a given subfamily, that sequence and a second highly homologous sequence (sometimes represented by a duplication of the first sequence) were aligned against every other sequence in the subfamily, one at a time, until every combination was generated. For example, three-way alignments were made between human ATR/*Xenopus* ATR and each of the other ATRs. The presence of two closely related sequences effectively "anchors" the alignment, precluding the patchwork alignment of short regions of fortuitously similar sequence that might otherwise occur; at the same time, inclusion of a more distantly related sequence forces the algorithm to detect the most significant homologies. The entire array of possible three-way alignments was generated twice, once with Clustal W and once with PSI-BLAST. For each combination of three sequences, the alignments obtained by the two algorithms were merged manually into a single three-way alignment. Any one three-way alignment, from either algorithm, shows 15%–25% identity amongst the compared sequences, when the alignments are merged this number is slightly higher.

The resulting collection of refined three-way alignments provides a large set of interrelated relationships among the different sequences that tightly constrain any composite all-inclusive alignment. Such alignments were constructed manually, with some assistance from PSI-BLAST and HNN. Closely related regions and sequences were aligned first; more distantly related sequences were then reconciled with one another through sequences that appear to be evolutionarily intermediate (e.g., linkage of ATR and *Neurospora* ATR via *E. nidulans* UVSB).

With the all-inclusive alignments in hand, the proteins of each subfamily were analyzed with respect to the possible presence of HEAT repeats. Identification of this motif in the absence of a 3D structure is problematic because the amino acid sequence signature is extremely flexible: the lengths and the amino acid compositions of the two helices and the intraunit loop can vary substantially from one unit to another (Andrade et al., 2001a, 2001b). Nonetheless, certain common features can be discerned from inspection of structurally defined HEAT repeat domains: helical regions are usually 10–20 residues, while intraunit loops tend to be 5–8 residues (e.g., Supplemental Figure S1 available at http://www.cell.com/cgi/content/full/112/2/151/DC1) (Andrade and

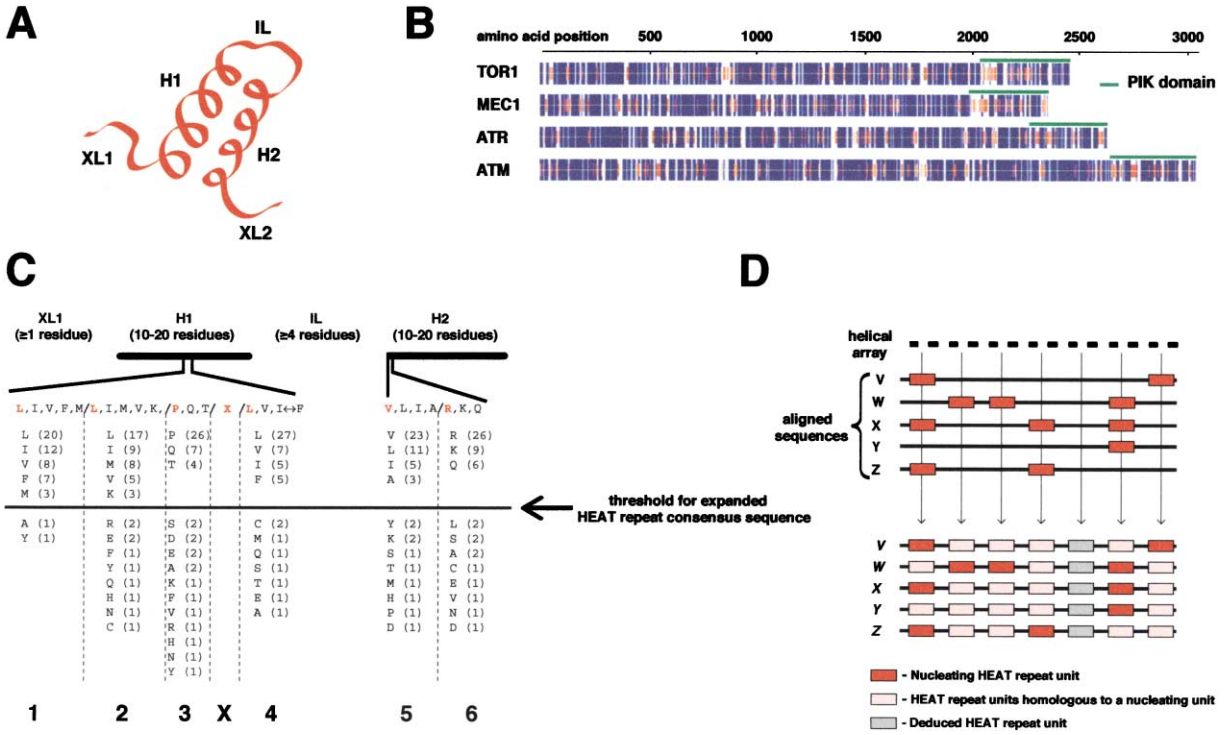

Figure 1. HEAT Repeat Identification Strategy

(A) General structure of a HEAT repeat unit. XL, external or interunit loop; IL, intraunit loop; H, helix.

(B) HNN 2° structure results. TOR1 and MEC1 are the *S. cerevisiae* homologs of mTOR and ATR, respectively. HEA*T* repeats were first identified in the N-terminal half of *T*OR1 (Andrade and Bork, 1995). Blue is α-helical, red is a mixture of extended strand and random coil structure.

(C) A structure based consensus sequence for identifying HEAT repeats (text).

(D) A strategy for using multiple sequence alignments to determine the location of highly divergent HEAT repeat units.

Bork, 1995; Cingolani et al., 1999; Chook and Blobel, 1999; Groves et al., 1999). A subset of HEAT repeat units is further defined by the presence of two specific amino acid sequence motifs, each of which is also a component of a specific structural feature: (1) a VR motif, which occurs at the N-terminal edge of the C-terminal helix; and (2) an LLPXL motif, which occurs within the N-terminal helix, usually near its central portion, and sometimes marks the site of a kink in the α-helical structure. These motifs were revealed by the presence of many such units in the A subunit of PP2A (Andrade and Bork, 1995). We have expanded the consensus sequence for these motifs slightly using information available from all three HEAT repeat proteins for which 3D structures are known (importin-β, karyopherin-β2, and the A subunit of PP2A, which together provide a set of 52 precisely defined HEAT repeat units; Cingolani et al., 1999; Chook and Blobel, 1999; Groves et al., 1999); (Figure 1C, Supplemental Figure S1 available at above website). Alignment of the structural equivalents of the VR and LLPXL motifs of these units permits the definition of an expanded consensus sequence consisting of any amino acid that occurs three or more times at each particular position (Figure 1C).

We have used this information to formulate rules that identify an expanded subset of HEAT repeats: (1) HNN or Jpred must predict a given segment to contain two helical patches with roughly appropriate spacing. (2) That segment must also contain sequence elements which match the expanded consensus sequence (above) at five of six, or six of six, positions, and (3) within the LLPXL motif, the first position must be a large hydrophobic residue, and cannot be a G, A, S, P, or T residue. The last constraint helps discriminate HEAT repeats from the evolutionarily related and structurally similar, but distinguishable, *arm*adillo (ARM) helical repeats (Supplemental Figure S2 available at above website; below). Any protein sequence segment that satisfies all three of these conditions was defined as a "nucleating HEAT repeat unit".

With these specific defining criteria in hand, each PIK-like protein subfamily was then analyzed for its total HEAT repeat content by a three-step approach (Figure 1C): (1) the comprehensive alignment, with accompanying secondary structure information, was examined for segments (in any of the component sequences) that adhere to the three criteria described above, thus defining all nucleating HEAT repeat units for that subfamily; (2) a nucleating unit at any particular location in one sequence implies the presence of an "homologous" HEAT repeat unit at the corresponding positions of all other sequences in the set, and thus highly divergent HEAT repeat units can be identified; (3) the definition used to identify nucleating units, while somewhat expanded as compared with previous criteria, is still quite particular; it identifies only 33/52 structurally defined HEAT repeat units from which it was derived (importin-β [10/19], karyopherin-β2 [10/18], and A-PP2A [13/15]).

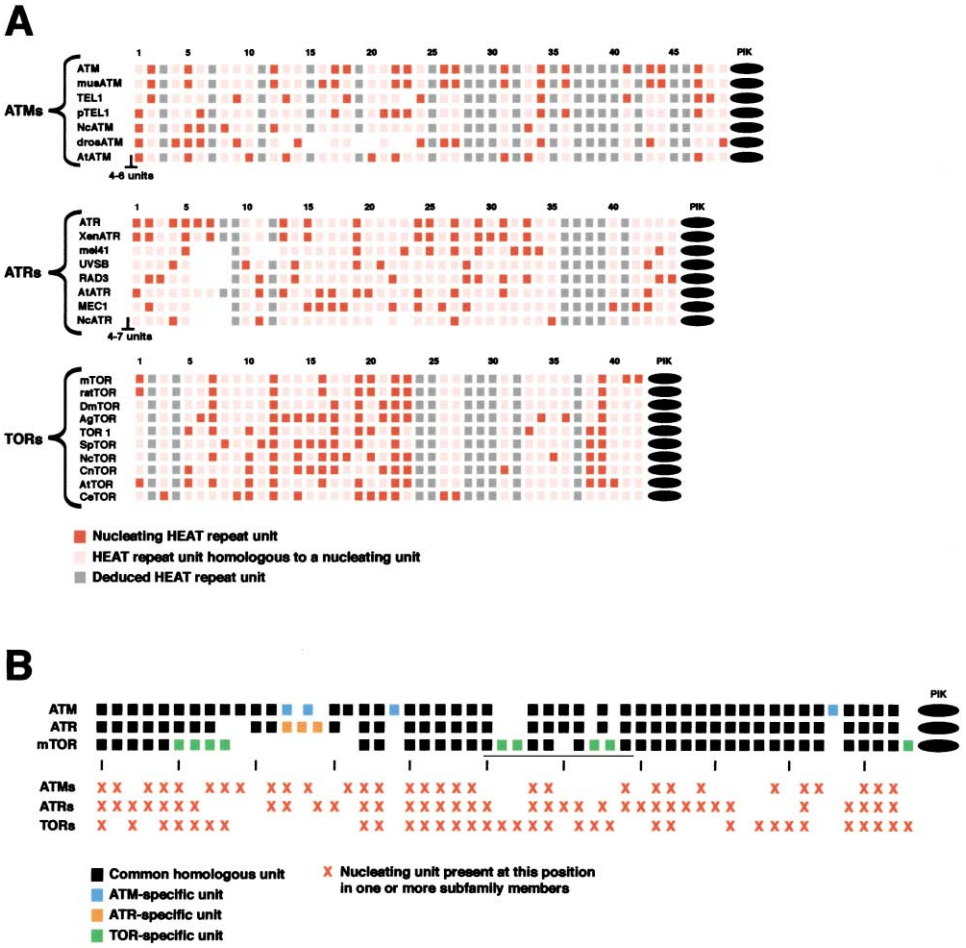

Figure 2. Analysis of ATRs, ATMs, and TORs

(A) HEAT repeat arrays of members of the ATR, ATM, and TOR PIK-like subfamilies.

(B) ATRs versus ATMs and TORs. The alignment of the HEAT repeat arrays of the three human proteins is shown. Note that at the unit level, ATR appears to be an evolutionary intermediate between ATM and mTOR.

Thus, all positions that did not emerge as "nucleating" or homologous units were examined for HEAT repeat potential. Any segment that (1) contained two adjacent helices of roughly the appropriate dimensions and (2) exhibited acceptable amino acids at consensus sequence positions, was tentatively assigned (pending further analysis; below) as a "deduced" HEAT repeat unit.

The results of this analysis are summarized in Figure 2A. All ATRs, ATMs, and TORs contain multiple nucleating HEAT repeat units, which occur in a unique patchwork for each sequence, and throughout the aligned sequences of each subfamily (red boxes). A total of 63%–83% of the non-kinase-domain residues in each analyzed subfamily are accounted for as nucleating or homologous-to-nucleating HEAT repeats (red plus pink boxes). Most remaining non-kinase-domain residues are accounted for as deduced HEAT repeats (gray boxes). The actual alignments and HEAT repeat assignments are provided as Supplemental Figures S3–S5 available at http://www.cell.com/cgi/content/full/112/2/151/DC1.

The above analysis also permits alignment of ATRs, ATMs, and TORs to one another. This was accomplished manually with some assistance from PSI-BLAST (Figure 2B; Supplemental Figures S6 and S7 available at above website). The resulting multisubfamily alignment provides three additional results. First, the basic HEAT repeat arrays of all three subfamilies are very closely related. The majority of identified HEAT repeat units are present in nearly all members of all three subfamilies; several more are present in nearly all members of two of the three subfamilies (mapped onto the sequences of the human proteins in Figure 2B, black boxes). Second, a few key units are "subfamily-specific", i.e., they are present in all (or nearly all) members of a single subfamily and are absent in all members of the other two subfamilies (Figure 2B, colored boxes). Third, a nucleating unit occurs in at least one member of at least one subfamily at virtually every position (summary in Figure 2B, bottom), providing additional evidence that all of these proteins are composed of HEAT repeats throughout their lengths and confirming the HEAT repeat nature of the "deduced" units.

To further support our HEAT repeat assignments, we carried out two complementary negative control studies.

First, the above conclusions rest on the assumption that the criteria used to define nucleating HEAT repeat

units identify HEAT repeat units specifically, rather than (also) identifying other types of helical repeat motifs or simply regions of high α helicity. We therefore applied our nucleating HEAT repeat constraints to two groups of ARM repeat proteins, which not only have high α-helical content but are composed of another type of iterated helical repeat unit that is structurally and evolutionarily related to a HEAT repeat (Andrade et al., 2001b). We inspected an alignment of ten *arm*adillo sequences, each containing 14 ARM repeats, and an alignment of seven importin-α sequences, each containing 10 ARM repeat units (Supplemental Figures S8 and S9 available at above website) (Huber et al., 1997; Conti et al., 1998). For armadillo, seven of 140 inspected ARM units are identified falsely as HEAT repeats and these units are confined to only two of 14 positions; for importin-α, three of seventy inspected units are misidentified, representing only two of ten positions. The set of "misidentified" units in the ARM repeat proteins is much more restricted than the broad array of segments identified as nucleating HEAT repeat units in the PIK-like proteins. The overall frequency of identified units was 5% (10/210) in the ARM proteins and 25% (274/1096) in the PIK-like proteins. More importantly, the misidentified ARM units are concentrated in only 17% of ARM repeat positions (4/24) while among the PIK-like proteins, an average of 75% of all positions are represented by at least one nucleating HEAT repeat unit as revealed by analysis of individual subfamilies (63% ATMs, 83% ATRs, 79% TORs), and >90% of positions were thus represented when all three subfamilies are considered with respect to one another (90% ATMs, 93% ATRs, 98% TORs; Figure 2B). This is the predicted pattern: HEAT repeats are much more diverse in sequence than ARM repeats, implying that at each position a smaller fraction of units will be detected as "nucleating". It also important to note that all of the misidentified segments among the ARM repeat proteins were in fact ARM repeat units, not random sequence segments. Thus, our criteria for nucleating HEAT repeat units are not simply identifying sequences that would be found in any protein of high α helicity.

These comparisons also reveal that the ARM repeats of importin-α are more closely related to HEAT repeats than are those of armadillo. With respect to the VR and LLPXL structural motifs alone (i.e., without consideration of 2° structure), 22 of 70 importin-α ARM repeats conform to the consensus defined for HEAT repeats as compared with 20 of 140 armadillo ARM repeats. This is primarily because the ARM repeats of importin-α, like HEAT repeats, frequently contain a large hydrophobic residue at the first position of the LLPXL motif.

Second, since ARM repeats are so closely related to HEAT repeats, and since some ARM repeat units would in fact be identified as nucleating HEAT repeat units by our analysis (above), we assessed directly the possible presence of ARM repeats within the PIK-like proteins. Application of an armadillo-based ARM repeat consensus (Supplemental Figure S2 available at http://www.cell.com/cgi/content/full/112/2/151/DC1, the major factor being a G, A, S, T, or P at the first position of the LLPXL motif for ARM repeats) to the ATRs reveals that only 12/341 identified HEAT repeats meet the minimum sequence requirements for a nucleating ARM repeat unit, and these occur at only 8 of the 45 ATR HEAT repeat positions (Supplemental Figure S10 available at above website). All twelve identifications are almost certainly spurious, for two reasons. First, many of these units exhibit additional sequence features, not included in the criteria for a nucleating unit (such as a large hydrophobic residue in the position before the LLPXL motif), which point to their being HEAT repeats rather than ARM repeats (Supplemental Figure S10 available at above website). Second, each of the corresponding positions is represented by a nucleating HEAT repeat unit in some other sequence(s) (Figure 2B, bottom). Analogously, in the HEAT repeat protein importin-β, one of 18 units might be misidentified as an ARM repeat (unit I10, Supplemental Figure S1 available at above website), a frequency statistically in accord our results for the ATRs. Nonetheless, we leave open the possibility that some misidentified units might represent interesting evolutionary intermediates between HEAT and ARM repeats.

The results described above suggest several general insights into the structure, evolution, and organization of the three analyzed PIK-like subfamilies. (1) ATRs, ATMs, and TORs have analogous structures dominated by massive N-terminal HEAT repeat domains followed by relatively small kinase domains. The HEAT repeat domains, which range in size from 40 to ~54 units, are comparable in size to those predicted for Huntingtin and GCN1 (Andrade and Bork, 1995). By analogy with structurally analyzed HEAT repeat domains, these regions are predicted to adopt large superhelical conformations that partially encompass their target macromolecules (above). This is consistent with recent results with the ATR and mTOR interacting molecules, ATRIP and RAPTOR, respectively, with which the appropriate PIK-like molecules associate weakly at many different positions (Cortez et al., 2001; Kim et al., 2002). (2) The three analyzed subfamilies are homologous to one another, not only in vicinity of their kinase domains, but throughout their entire lengths. (3) Different sequences have diverged from one another by "modular" changes, as would be expected for proteins of iterative structure. Most of these changes involve the addition/subtraction of one or a few HEAT repeat units, and many are "subfamily specific". There is also an interesting possibility of a modular translocation (Supplemental Figure S7 available at above website). The subfamily specific units occur primarily in two regions, one near the N terminus and the other immediately upstream of the kinase domain (Figure 2B). One particularly striking example is provided by the rapamycin binding domain (RBD) of the TORs, which is located immediately adjacent to the conserved kinase domain. Previously described as a four-helix bundle, the RBD emerges here as a pair of HEAT repeat units, one of which is TOR-specific (Supplemental Figure S7 available at above website) (Choi et al., 1996).

The above results provide a framework for further investigation of the roles and mechanisms of ATRs, ATMs, and TORs. The general conclusions described for these three groups of proteins appear to apply to all other proteins of the PI3K-like superfamily (DNA-PKs, SMG-1s, and TRRAPs), as will be described elsewhere. The presence of a common, evolutionarily conserved overall structure among all PI3K-like proteins raises the possibility that, despite their diverse biological roles, all

of these proteins share common underlying properties in their basic biochemical mechanisms of action.

Acknowledgments

J. P. was supported by N.I.H. grant RO1-GM 44794 to N.K. We thank Drs. Lewis Cantley, Job Dekker, Stuart Schreiber, and Beth Weiner for insightful discussions and all contributors to the various complete genome sequencing projects, whose efforts have made this work possible. Sequence accession numbers are provided in the Supplemental Material available at available at http://www.cell.com/cgi/content/full/112/2/151/DC1.

Jason Perry* and Nancy Kleckner
Department of Molecular and Cellular Biology
Harvard University
Cambridge, Massachusetts 02138

References

Abraham, R.T. (2001). Genes Dev. *15*, 2177–2196.

Andrade, M.A., and Bork, P. (1995). Nat. Genet. *11*, 115–116.

Andrade, M.A., Perez-Iratxeta, C., and Ponting, C.P. (2001a). J. Struct. Biol. *134*, 117–131.

Andrade, M.A., Petosa, C., O'Donoghue, S.I., Muller, C.W., and Bork, P. (2001b). J. Mol. Biol. *309*, 1–18.

Cha, R.S., and Kleckner, N. (2002). Science *297*, 602–606.

Choi, J., Chen, J., Schreiber, S.L., and Clardy, J. (1996). Science *273*, 239–242.

Chook, Y.M., and Blobel, G. (1999). Nature *399*, 230–237.

Cingolani, G., Petosa, C., Weis, K., and Muller, C.W. (1999). Nature *399*, 221–229.

Conti, E., Uy, M., Leighton, L., Blobel, G., and Kuriyan, J. (1998). Cell *94*, 193–204.

Cortez, D., Guntuku, S., Qin, J., and Elledge, S.J. (2001). Science *294*, 1713–1716.

Grant, P.A., Schieltz, D., Pray-Grant, M.G., Yates, J.R., and Workman, J.L. (1998). Mol. Cell *2*, 863–867.

Groves, M.R., Hanlon, N., Turowski, P., Hemmings, B.A., and Barford, D. (1999). Cell *96*, 99–110.

Huber, A.H., Nelson, W.J., and Weis, W.I. (1997). Cell *90*, 871–882.

Kim, D.H., Sarbassov, D.D., Ali, S.M., King, J.E., Latek, R.R., Erdjument-Bromage, H., Tempst, P., and Sabatini, D.M. (2002). Cell *110*, 163–175.

Ma, Y., Pannicke, U., Schwarz, K., and Lieber, M.R. (2002). Cell *108*, 781–794.

McMahon, S.B., Buskirk, H.A.V., Dugan, K.A., Copeland, T.D., and Cole, M.D. (1998). Cell *94*, 363–374.

Rosen, E.M., Fan, S., Goldberg, I.D., and Rockwell, S. (2000). Oncology *14*, 741–757.

Schmeizle, T., and Hall, M.N. (2000). Cell *103*, 253–262.

Shiloh, Y., and Kastan, M.B. (2001). Adv. Cancer Res. *83*, 209–254.

Vassiley, A., Yamauchi, J., Kotani, T., Prives, C., Avantaggiati, M.L., Qin, J., and Nakatani, Y. (1998). Mol. Cell *2*, 869–875.

Yamashita, A., Ohnishi, T., Kashima, I., Taya, Y., and Ohno, S. (2001). Genes Dev. *15*, 2215–2228.

*Correspondence: jperry@fas.harvard.edu